Sequence modeling and design from molecular to genome scale with Evo

Year: 2024 | Citations: 212 | Authors: Eric Nguyen, Michael Poli, Matthew G. Durrant, Armin W. Thomas, Brian Kang

Abstract

The genome is a sequence that completely encodes the DNA, RNA, and proteins that orchestrate the function of a whole organism. Advances in machine learning combined with massive datasets of whole genomes could enable a biological foundation model that accelerates the mechanistic understanding and generative design of complex molecular interactions. We report Evo, a genomic foundation model that enables prediction and generation tasks from the molecular to genome scale. Using an architecture based on advances in deep signal processing, we scale Evo to 7 billion parameters with a context length of 131 kilobases (kb) at single-nucleotide, byte resolution. Trained on 2.7M prokaryotic and phage genomes, Evo can generalize across the three fundamental modalities of the central dogma of molecular biology to perform zero-shot function prediction that is competitive with, or outperforms, leading domain-specific language models. Evo also excels at multi-element generation tasks, which we demonstrate by generating synthetic CRISPR-Cas molecular complexes and entire transposable systems for the first time. Using information learned over whole genomes, Evo can also predict gene essentiality at nucleotide resolution and can generate coding-rich sequences up to 650 kb in length, orders of magnitude longer than previous methods. Advances in multi-modal and multiscale learning with Evo provides a promising path toward improving our understanding and control of biology across multiple levels of complexity.